DTIC FILE COPY

AD-A226 869

REPORT DOCUMENTATION PAGE

1a.	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY NA	3. DISTRIBUTION / AVAILABILITY OF REPORT Approved for public release; distribution unlimited
2b. DECLASSIFICATION / DOWNGRADING SCHEDULE NA	
4. PERFORMING ORGANIZATION REPORT NUMBER(S)	5. MONITORING ORGANIZATION REPORT NUMBER(S) AFOSR-TR- 90 1034
6a. NAME OF PERFORMING ORGANIZATION University of Texas at Austin	6b. OFFICE SYMBOL (If applicable)
7a. NAME OF MONITORING ORGANIZATION AFOSR/NM	
6c. ADDRESS (City, State, and ZIP Code) Department of Electrical and Computer Engr. Austin, Texas 78712-1024	7b. ADDRESS (City, State, and ZIP Code) Bldg. 410 Bolling AFB, DC 20332-6448
8a. NAME OF FUNDING / SPONSORING ORGANIZATION AFOSR	8b. OFFICE SYMBOL (If applicable) NM
9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER AFOSR-86-0026	
8c. ADDRESS (City, State, and ZIP Code) Bldg. 410 Bolling AFB, DC 20332-6448	10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT ACCESSION NO.
6.1102F	2304	A6	

11. TITLE (Include Security Classification)
SOME APPLICATIONS OF PROBABILITY AND STATISTICS IN COMMUNICATION THEORY AND SIGNAL PROCESSING

12. PERSONAL AUTHOR(S)
Gary L. Wise

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Technical Final	FROM 11/1/86 TO 4/30/90	1990, August, 16	21

16. SUPPLEMENTARY NOTATION

17. COSATI CODES

FIELD	GROUP	SUB-GROUP

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
Estimation theory, martingale convergence theorem,

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

This Final Technical Report constitutes a summary of the research performed under Grant AFOSR-86-0026 during the period November 1, 1986 through April 30, 1990. First we present a list of the personnel involved in the research effort. Then in the following section we present a brief summary of the research results that have been achieved. Each of these results is well documented in technical articles, and references to these articles are made in the summary of the research results.

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT	21. ABSTRACT SECURITY CLASSIFICATION
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	UNCLASSIFIED

22a. NAME OF RESPONSIBLE INDIVIDUAL	22b. TELEPHONE (Include Area Code)	22c. OFFICE SYMBOL
Dr. J. Sjogren	(202) 767-5027	AFOSR/NM

DD FORM 1473, 84 MAR   83 APR edition may be used until exhausted. All other editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED

FINAL TECHNICAL REPORT

GRANT AFOSR-86-0026

SOME APPLICATIONS OF PROBABILITY AND STATISTICS IN COMMUNICATION THEORY AND SIGNAL PROCESSING

Air Force Office of Scientific Research
Attention: Dr. J. Sjogren
Bolling Air Force Base
Washington, DC 20332-6448

Gary L. Wise
Department of Electrical and Computer Engineering
and
Department of Mathematics
The University of Texas at Austin
Austin, Texas 78712

INTRODUCTION

This Final Technical Report constitutes a summary of the research performed under Grant AFOSR-86-0026 during the period November 1, 1986 through April 30, 1990. First we present a list of the personnel involved in the research effort. Then in the following section we present a brief summary of the research results that have been achieved. Each of these results is well documented in technical articles, and references to these articles are made in the summary of the research results.

PERSONNEL

Principal Investigator:

Gary L. Wise

Postdoctoral Research Associates:

John M. Morrison

Alan E. Wessel

Graduate Research Associate:

Eric B. Hall

A SURVEY OF RESULTS

In this final technical report we will briefly comment upon our research accomplishments sponsored by the Grant AFOSR-86-026. Much of our work during this period was concerned with various aspects of estimation theory. Additional work was done in the areas of contention resolution for local area computer networks, signal detection theory, data compression for image processing, and linear system theory.

Typically, the problem of estimation is concerned with attempting to approximate a desired quantity by a function of the available data so as to minimize a prescribed fidelity criterion. A commonplace example might be given by attempting to estimate a second order random variable X (perhaps a signal of interest) by some function $f(\cdot)$ of the datum Y (perhaps a noise corrupted version of the signal) so as to minimize the mean square error $E([X - f(Y)]^2)$. This example appears in many works on the subject of estimation theory. In earlier work, sponsored by a previous AFOSR Grant, we showed [1] that the best such function is not necessarily given by $f(Y) = E(X \mid Y)$, even thought X and Y are both bounded random variables. Moreover, it might seem that there is little justification from a practical viewpoint of choosing the mean square error as the appropriate fidelity criterion. Consider a fidelity criterion given by the expectation of a cost function of the error. In the context of estimation theory, one is often confronted with two concerns in choosing a cost function: the concern that the cost function adequately reflects the cost one wishes to attach to an error, and the concern that the cost function results in a problem which one finds to be mathematically tractable. A cursory inspection of the literature in estimation theory might suggest that in many cases the second of the above concerns totally eclipses the first concern. We began a serious study of estimation theory. This work was directed to the very underpinnings of estimation theory, and it is representative of what in many cases in the literature is ignored, is postulated with no concern for the consistency of everything being postulated, or is otherwise swept under the rug. The two such areas in which we have achieved some success are concerned with continuity properties of filtrations of σ-algebras generated by stochastic processes and with the convergence rate of the martingale convergence theorem. We will now briefly comment on our results in these areas.

Let $(\Omega, \mathcal{F}, P)$ be a probability space. We take a filtration of σ-algebras to be any nondecreasing collection of σ-subalgebras of $\mathcal{F}$ indexed by $[0, \infty)$. Let $\{\mathcal{F}_t : t \geq 0\}$ be a filtration. Define $\mathcal{F}_{0-} = \mathcal{F}_0$; otherwise, define $\mathcal{F}_{t-} = \bigvee_{s<t} \mathcal{F}_s$ and $\mathcal{F}_{t+} = \bigcap_{s>t} \mathcal{F}_s$.

We say that a filtration is continuous at t if $\mathcal{F}_{t-} = \mathcal{F}_{t+}$; we say that it is left continuous at t if $\mathcal{F}_{t-} = \mathcal{F}_t$; and we say that it is right continuous at t if $\mathcal{F}_t = \mathcal{F}_{t+}$. The filtration $\{\mathcal{F}_t : t \geq 0\}$ is continuous, left continuous, or right continuous if it is continuous, left continuous, or right continuous, respectively, at t for all $t \geq 0$. Let $\{X(t): t \in [0, \infty)\}$ be a stochastic process defined on $(\Omega, \mathcal{F}, P)$. By a P-null set we mean a measurable set which has probability zero. The canonical filtration of this stochastic process is given by $\mathcal{F}_t = \sigma(X(s): s \leq t) \vee (\text{P-null sets})$ for $t \geq 0$.

In the context of estimation theory where the data are represented by a stochastic process indexed by an interval of real numbers, much of the current literature is concerned with stochastic differential equations and with martingale theory. Stochastic differential equations often arise as models for stochastic dynamical systems and techniques from martingale theory often arise in the analysis of estimation schemes and their approximation properties. In these areas one often encounters hypotheses stipulating the right continuity of filtrations of σ-algebras generated by stochastic processes. This is a blanket assumption made by many in the French and Soviet schools of stochastic process theory; see, for example, [2], [3], [4], and [5]. However, the question emerges as to when this assumption is justified or as to what reasonable hypotheses might imply it. It is often tempting and pleasing to the intuition to believe that the regularity of the sample paths of a stochastic process and the continuity of its associated canonical filtration are closely related. For example, separable Brownian motion has almost surely continuous sample paths and with the aid of the Blumenthal Zero-One Law [6] we see that its canonical filtration is continuous. Conversely, martingales with respect to right continuous filtrations have versions that are almost surely cadlag [7]. If we heuristically think of the canonical filtration $\mathcal{F}_t$ as the "data" conveyed by the stochastic process $\{X(t): t \in [0, \infty)\}$ up to and including time t, we may be inclined to suppose that the continuity of the sample paths of the process might prevent jumps in the "data" $\{\mathcal{F}_t : t \geq 0\}$; and we also might suppose that the continuity of the "data" flow would influence the regularity of the sample paths of the stochastic process. (In [1] we pointed out that this is a totally erroneous concept of data.) In [8] we investigated what properties characterize filtrations of σ-algebras that are continuous. In this work we showed that the regularity of the sample paths of a stochastic process and the continuity of its associated filtration are logically independent; we presented an example of a stochastic process with infinitely differentiable sample paths and a discontinuous canonical filtration and we also gave an example where a stochastic process could have an arbitrarily irregularly prescribed sample path (e.g. non-Lebesgue measurable) and a continuous canonical filtration. We also presented an example of a stochastic process whose canonical filtration was discontinuous at every point. We then went on and established conditions guaranteeing the continuity of a filtration of σ-algebras. Also, we presented necessary and sufficient conditions for a filtration of σ-algebras to be continuous, right continuous, or left continuous. For example, we established the following results in [8]:

Theorem: Let $(\Omega, \mathcal{F}, P)$ be a probability space and $\{\mathcal{F}_t : t \geq 0\}$ be a filtration on $(\Omega, \mathcal{F}, P)$ so that $\mathcal{F}_0$ contains the P-null sets. Then the filtration is continuous at t_0 if and only if for all $Y \in L_2(\Omega, \mathcal{F}, P)$, the stochastic process defined by $Y_t = E(Y \mid \mathcal{F}_t)$, $t \geq 0$, is L_2 continuous at t_0.

Let $(\Omega, \mathcal{F}, P)$ be a probability space. A σ-subalgebra $\mathcal{A}$ of $\mathcal{F}$ is said to be essentially countably generated if there exists a countable subset K of $\mathcal{F}$ so that $\sigma(K) \vee (\text{P-null sets}) = \mathcal{A} \vee (\text{P-null sets})$.

Theorem: Let M be a separable metric space and $\{X(t): t \geq 0\}$ be a stochastic process taking values in M that is left or right continuous in probability. Then $\sigma(X(t): t \geq 0)$ is essentially countably generated.

Theorem: Let $(\Omega, \mathcal{F}, P)$ be a separable probability space. Then if $\{\mathcal{F}_t : t \geq 0\}$ is a filtration on $(\Omega, \mathcal{F}, P)$ so that $\mathcal{F}_0$ contains the P-null sets, there exists a countable subset C of $\mathbb{R}$ so that for $t \notin C$, $\mathcal{F}_{t-} = \mathcal{F}_{t+} = \mathcal{F}_t$.

Now we comment upon some of our recent results on martingales. Frequently, in estimation theory one derives a sequence of estimators, say Y_n, and one desires to show that as $n \to \infty$, Y_n converges in an appropriate sense. A typical example arises in an attempt to estimate a second order random variable X as a function of the available data, say $\{Z_n: n \in \mathbb{N}\}$, by choosing $Y_n = E(X \mid Z_1, Z_2, \ldots, Z_n)$. In this endeavor, the martingale convergence theorem often surfaces as a useful tool in establishing convergence. However, in a practical circumstance, if one were interested in convergence and if n corresponded to the progression of time, then the rate of convergence would also be of concern. This would arise, for instance, if Y_n represented the estimate after n samples of data are taken and data is sampled at regularly spaced intervals. The key to establishing this rate of convergence is intimately linked with the convergence rate of the martingale convergence theorem. In [9] we examined the convergence rate of the martingale convergence theorem, and we showed that this convergence can be nonuniform and, consequently, arbitrarily slow. This result that the convergence rate of the martingale convergence theorem can be arbitrarily slow is important not only from the obvious practical viewpoint, but also from the viewpoint of the mathematician, since the martingale convergence theorem is one of the key theorems of probability theory.

Another aspect of the martingale convergence theorem which we investigated was concerned with the use of the martingale convergence theorem in estimating a random variable X. Let X be a second order random variable, and let $\{Z_n: n \in N\}$ be a sequence of random variables representing data. Often one may attempt to estimate X based upon the first n terms of the data sequence by $E(X \mid Z_1, Z_2, \ldots, Z_n)$. In [10] we pointed out some problems associated with an overly cavalier usage of the martingale convergence theorem in this context. In particular, we gave an example where each of the above random variables was zero mean Gaussian with a positive variance, $E(X \mid Z_1, Z_2, \ldots, Z_n) = 0$ almost surely for each $n \in N$, and yet for any positive integer k there exists a function $f_k:\mathbf{R}\rightarrow\mathbf{R}$ so that $f_k(Y_k) = X$ pointwise on the underlying probability space.

In a similar context as the above, in [11] we noted that for a second order random variable X, the rate of the L_2 convergence of $E[X \mid Y_1, Y_2, \ldots, Y_n]$ can crucially depend upon X. That is, any L_2 perturbation in X could drastically alter the rate of convergence.

Another aspect of estimation theory with which we were concerned dealt with the idea of when an estimator which was optimal under a given fidelity criterion would also be optimal under certain other fidelity criteria. A classical paper on this subject in [12] was written by Sherman, and this result is known in the engineering literature as Sherman's theorem. However, a close inspection of [12] shows some erroneous claims. In [13] we presented a correct derivation of the effort undertaken in [12]. The following theorem is a correct version of Sherman's theorem and we proved it in [13].

Theorem: Let $k \in \mathbf{N}$, $(\Omega, \mathcal{S}, P)$ be a probability space, and $X, Y_1, \ldots, Y_k$ be random variables defined on $(\Omega, \mathcal{S}, P)$, with X integrable. Let $M:\mathbf{R}^k \rightarrow \mathbf{R}$ be a Borel measurable function such that $M[Y_1(\omega), \ldots, Y_k(\omega)] = E[X \mid Y_1, \ldots, Y_k](\omega)$ a.s., and assume that there exists a regular conditional distribution function of X conditioned on $\sigma(Y_1, \ldots, Y_k)$, $F:\mathbf{R} \times \Omega \rightarrow [0,1]$, such that $F(x+M[Y_1(\omega), \ldots, Y_k(\omega)],\omega)$, as a function of x with ω fixed, is unimodal about the origin and symmetric. Then $M[Y_1, \ldots, Y_k]$ minimizes the quantity $E[\Phi(X-f(Y_1,\ldots,Y_k))]$ over all Borel measurable functions $f:R^k \to R$ where $\Phi:R \to [0,\infty)$ is even and nondecreasing on $[0,\infty)$.

Several attempts at a proof of the above result have been presented in the engineering literature, and each that we know of is wrong; counterexamples to these efforts are given in [14].

Thus, the result in the above theorem requires a regular conditional distribution function that, when properly shifted, is symmetric and unimodal about the origin and a cost function that is nonnegative, even, and nondecreasing to the right of the origin. It is easy to see that if in this theorem we let k=1 and X and Y be mutually Gaussian random variables then the resulting regular conditional distribution function is symmetric and unimodal about $E[X|Y](\omega)$ for any fixed ω. This special case explains why Sherman's theorem is often invoked to add a token claim of generality to papers that only consider Gaussian distributions. When one attempts to venture outside this somewhat limited arena, however, the conditions which Theorem 1 places on the regular conditional distribution function immediately begin to feel overly restrictive. After all, how comfortable should we be with the assumption that the regular conditional distribution function under consideration is unimodal about the conditional mean? The conditions on the cost function, on the other hand, are extremely nonrestrictive and, in fact, allow for many interesting, albeit impractical, choices. For example, the cost function given by

$$\Phi(x) = \int_0^{|x|} I_C(t)\,dt,$$

where C denotes a Cantor subset of $[0,\infty)$ of positive Lebesgue measure, satisfies the conditions of the above theorem. This imbalance suggests the possibility of lessening the restrictions on the regular conditional distribution function by perhaps slightly increasing the restrictions imposed on the cost function. In [14], we presented a more general treatment of this general subject. The following results are presented in [14]. Notice that the first result dispenses with the unimodality assumption, and the second result allows us to base our estimate upon random variables measurable with respect to a non countably generated σ-algebra, such as, for instance, that which may be generated by a random object.

Theorem: Let $k \in N$, $(\Omega, \mathcal{S}, P)$ be a probability space, and $X, Y_1,\ldots,Y_k$ be random variables defined on $(\Omega, \mathcal{S}, P)$, with X integrable. Let $M:R^k \to R$ be a Borel measurable function such that $M[Y_1(\omega),\ldots,Y_k(\omega)] = E[X \mid Y_1,\ldots,Y_k](\omega)$ a.s., and assume that there exists a regular conditional distribution function of X conditioned on $\sigma(Y_1,\ldots,Y_k)$,

$F:\mathbf{R} \times \Omega \to [0,1]$, such that $F(x+M[Y_1(\omega), \ldots, Y_k(\omega)],\omega)$, as a function of x with ω fixed, is symmetric. Then $M[Y_1, \ldots, Y_k]$ minimizes the quantity $E[\Phi(X-f(Y_1, \ldots, Y_k))]$ over all Borel measurable functions $f:\mathbb{R}^k \to \mathbb{R}$ where $\Phi:\mathbb{R} \to [0, \infty)$ is even and convex.

Theorem: Let $(\Omega, \mathcal{S}, P)$ be a probability space, $\mathcal{A}$ be a σ–subalgebra of $\mathcal{S}$, and X be a random variable defined on $(\Omega, \mathcal{S}, P)$ such that X is integrable. For each $\omega \in \Omega$, let $M(\omega) = E[X|\mathcal{A}](\omega)$, and assume that there exists a regular conditional distribution function of X conditioned on $\mathcal{A}$, $F:\mathbb{R} \times \Omega \to [0,1]$, such that $F(x+M(\omega),\omega)$, as a function of x with ω fixed, is symmetric. Then M minimizes the quantity $E[\Phi(X-\hat{X})]$ over all $\mathcal{A}$–measurable random variables $\hat{X}$, where $\Phi:\mathbb{R} \to [0, \infty)$ is even and convex.

In [15] and [16] our concern was directed toward fusing, or combining, estimates based upon a finite number of estimates of a fixed second order random variable X in order to achieve a single "best" estimate of X. For example, if $X, Y_1, Y_2, \ldots, Y_n$ are random variables and X is second order, how might $E[X \mid Y_1], E[X \mid Y_2], \ldots, E[X \mid Y_n]$ be combined so as to approximate X in a minimum mean square sense? Although aspects of this problem have been considered in the literature, we know of no other work in this area that is correct. To illustrate some subtleties in this area, note the following two examples.

Example: For an integer $n > 1$, consider a set of random variables $\{X, Y_1, \ldots, Y_n\}$ with a joint probability density function given by

$$f(x, y_1, \ldots, y_n) = \left(\frac{1}{\sqrt{2\pi}}\right)^{n+1} \exp\left[\frac{-1}{2}\left(x^2 + \sum_{i=1}^{n} y_i^2\right)\right] \left[1 + x \exp\left(\frac{-x^2}{2}\right) \prod_{i=1}^{n} \left(y_i \exp\left(\frac{-y_i^2}{2}\right)\right)\right].$$

It follows straightforwardly that the set $\{X, Y_1, \ldots, Y_n\}$ is not mutually Gaussian and not mutually independent, yet any proper subset of $\{X, Y_1, \ldots, Y_n\}$ containing at least two random variables is mutually independent, mutually Gaussian, and identically distributed with each random variable having zero mean and unit variance. For any nonempty proper subset $\mathcal{D}$ of $\{Y_1, \ldots, Y_n\}$, we note that $E[X \mid \mathcal{D}] = 0$ a.s. since X is independent of $\mathcal{D}$. However, it follows quickly that

$$E[X \mid Y_1, \ldots, Y_n] = \frac{1}{2\sqrt{2}} Y_1 \cdots Y_n \exp\left[\frac{-1}{2}\left(Y_1^2 + Y_2^2 + \ldots + Y_n^2\right)\right] \text{ a.s.}$$

Thus, since any Borel measurable function of the estimates $E[X \mid \mathcal{D}]$ where $\mathcal{D}$ ranges over all nonempty proper subsets of $\{Y_1, \ldots, Y_n\}$ would be constant almost surely, it would not be reasonable to attempt to estimate $E[X \mid Y_1, \ldots, Y_n]$ based on a combination of these estimates.

Example: Let $\Omega = [0, 1]$, $\mathcal{F}$ denote the Borel subsets of Ω, and P denote Lebesgue measure on $\mathcal{F}$. Let A be a positive real number, $\sigma(Y_1) = \sigma([0, 1/2))$, $\sigma(Y_2) = \sigma([1/4, 3/4))$, and $X(\omega) = A$ for $\omega \in [0, 1/4) \cup [1/2, 3/4)$ and $X(\omega) = -A$ for $\omega \in [1/4, 1/2) \cup [3/4, 1]$. Then it straightforwardly follows that $E[X|Y_1] = E[X|Y_2] = 0$ a.s., but $E[X|Y_1, Y_2] = X$ a.s. Notice that in this special case, any linear combination of $E[X|Y_1]$ and $E[X|Y_2]$ yields an estimate equal to 0 a.s., resulting in a mean square error in approximating X of A^2, which can exceed any preassigned real number. Recalling that $E[X|Y_1]$ and $E[X|Y_2]$, respectively, are $\sigma(Y_1)$-measurable and $\sigma(Y_2)$-measurable, we see that $E[X|Y_1] = E[X|Y_2] = 0$ pointwise in ω; similarly, we see that $E[X|Y_1, Y_2] = X$ pointwise in ω. Thus, in this situation, it is fruitless to attempt to approximate X based on any function of $E[X|Y_1]$ and $E[X|Y_2]$.

In [15] and [16] we proved the following theorem.

Theorem: Consider a probability space $(\Omega, \mathcal{F}, P)$ and random variables $X, N_1, \ldots, N_n$ defined on $(\Omega, \mathcal{F}, P)$ where n is a positive integer and X is a second order random variable. Further, assume that for each positive integer $i \leq n$, N_i has a zero mean Gaussian distribution with positive variance given by σ_i^2, *and that* $X, N_1, \ldots, N_n$ *are mutually independent.* Define $Y_i = X + N_i$ for $i = 1, \ldots, n$. Then there exists a Borel measurable function $g: \mathbf{R}^n \rightarrow \mathbf{R}$ such that $E[X|Y_1, \ldots, Y_n] = g(E[X|Y_1], \ldots, E[X|Y_n])$ a.s.

A Monte Carlo variance reduction technique known as importance sampling has recently been applied to many problems in data communications. This technique holds the promise of offering vast improvements to traditional Monte Carlo methods. In [17] and [18] we considered importance sampling applied to the estimation of tail probabilities. In this work we gave counterexamples to some commonly used types of importance sampling. Then we introduced a new method of importance sampling, which we called Importance Sampling via a Simulacrum, and we illustrated how it could outperform some other methods of importance sampling.

In other papers we pointed out how wrong certain commonly accepted techniques and results in statistical signal processing can be. In [19] we presented a collection of counterexamples in detection and estimation. In [20] we presented a collection of counterexamples in conditioning. In [21], we presented a collection of counterexamples in maximum likelihood estimation. In [22] and [23] we presented some comments on some problems in Kalman filtering. The papers noted in this paragraph provide several counterexamples to what is often taken as common knowledge in the literature of statistical signal processing. A copy of [20] is appended to this report.

Another direction of our research efforts was in the area of contention resolution for local area computer networks. In the last few years, packet broadcasting random multiple-access computer communication networks have been commercially available. A typical example of such a network is the Ethernet, developed by Xerox, which was designed based on the idea of carrier sense multiple access with collision detection. In Ethernet, a station among a number of users sharing a common channel will listen before transmitting and defer if the channel is busy; when two or more stations collide, each colliding station waits for a random period of time before retransmitting. Although Ethernet has the advantage of easy interconnection of stations to the common channel and it provides a high level of utilization of the channel, it does not truly address the problem of how to effectively resolve collisions in the channel. Thus, a packet involved in a collision may incur excessive delay due to waiting and abortion of transmission. Recently, a protocol called Enet II was introduced [24] as a candidate for the second generation of Ethernet. This protocol is designed to effectively resolve contention in a broadcast multiple-access network such as Ethernet. We investigated the Enet II protocol in [25], and in this investigation, we gave expressions for the average time required to resolve a collision involving k stations and the average time for a particular station involved in a k-way collision to send its packet successfully. Our results in this area were derived analytically, without recourse to efforts based on approximations or simulations. In [25] we also considered the efficiency of the protocol, and we derived a lower bound for the maximum efficiency.

In the area of image processing, a modest effort was directed toward studying the properties of a data compression scheme for image processing. In [26] we considered a modification of an existing data compression scheme which allowed more general ways of processing the image data while maintaining the favorable data compression rates.

We also devoted some effort to the problem of signal detection. In [27] we studied the problem of choosing the nonlinearity $g(\cdot)$ when the test statistic was constrained to be of the form

$$\sum_{i=1}^{n} g(x_i),$$

where the x_i's represented our observations. Observe that in the case of a constant signal additively corrupted by mutually independent, identically distributed noise, the Neyman-Pearson test statistic is of the above form. If the noise sequence were not mutually independent, then the test statistic would not necessarily be of this form. However, it might seem reasonable to suppose that in some cases, if the noise were "almost mutually independent" then a test statistic of the above form might be a reasonable approximation to an appropriate test statistic. In [27] we studied the problem of choosing the function $g(\cdot)$ so as to maximize the asymptotic relative efficiency of this detector relative to any other detector of this form with a different nonlinearity.

In [28] we studied another aspect of statistical hypothesis testing. Consider the situation of testing one simple hypothesis against another simple hypothesis. The likelihood ratio (i.e. a Radon-Nikodym derivative) often arises; and it is known that in several contexts (e.g. Neyman-Pearson, Bayes, minimax) an optimum test is given by comparing the likelihood ratio against an appropriately chosen threshold. In [28] we studied the question of when a likelihood ratio with respect to two probability measures P_0 and P_1 might also be the likelihood ratio with respect to another pair of probability measures Q_0 and Q_1 on the same measurable space. In this way, one likelihood ratio might implement an optimum processing of the data for many pairs of probability measures; that is, an optimal data processor might be optimal even when different probability measures are governing the data. For the moment, consider the case where P_0 is absolutely continuous with respect to P_1; we gave examples where the Radon-Nikodym derivative $\frac{dP_0}{dP_1}$ was the likelihood ratio not only for testing P_0 against P_1, but also for testing Q_0 against Q_1, even when P_0 was extremely dissimilar from Q_0 and P_1 was extremely dissimilar from Q_1.

In some recent efforts, we have investigated some aspects of linear systems. Although the subject of linear systems has truly matured as a research area, we have uncovered some unappreciated aspects of the theory. In [29] (see also [30]) we investigated the representation of linear systems. In this work we established the following result.

<u>Theorem:</u> Let Ω be a locally compact separable metric space, μ be a σ-finite measure on $\mathcal{B}(\Omega)$ (where we use $\mathcal{B}(\cdot)$ to denote the Borel subsets of a topological space), and λ be a Borel measure on a locally compact separable metric space W. Let $T:L^1_{loc}(\Omega, \mathcal{B}(\Omega), \mu) \to L^1_{loc}(W, \mathcal{B}(W), \lambda)$ be a positive, continuous, linear map. Then there exists $K:\mathcal{B}(W) \times \Omega \to [0, \infty]$ so that

(i) For each $\omega \in \Omega$, $K(\cdot, \omega)$ is a regular Borel measure on $\mathcal{B}(W)$.

(ii) For each $E \in \mathcal{B}(W)$, $K(E, \cdot)$ is measurable on Ω.

(iii) For each $A \in \mathcal{B}(\Omega)$ with $\mu(A) < \infty$, the measure

$$\int_A K(\cdot, \omega)\, d\mu(\omega)$$

defined on $\mathcal{B}(W)$ is regular and λ-absolutely continuous.

(iv) $$T(f) = \frac{d}{d\lambda} \int_\Omega K(\cdot, \omega) f^+(\omega)\, d\mu(\omega) - \frac{d}{d\lambda} \int_\Omega K(\cdot, \omega) f^-(\omega)\, d\mu(\omega)$$

for $f \in L^1(\Omega, \mathcal{B}(\Omega), \mu)$, where by this notation, we mean the difference of the Radon-Nikodym derivatives of the measures given by the integrals.

Convolution frequently arises in the study of linear systems. In [31] we constructed two bounded, Lebesgue integrable, nowhere zero functions whose convolution is identically zero. This phenomenon seems to have been overlooked by many working in the area of linear systems. In particular, it dashes any hope of deconvolution in this situation. Also, although it is well known that $L_1(R)$, equipped with the operations of pointwise addition and convolution, is a commutative Banach algebra, this result shows that this commutative Banach algebra $L_1(R)$ is not an integral domain. Indeed, it shows much more than this, since there exist two *nowhere* zero integrable functions whose convolution is everywhere zero. In [32] we showed the analogous result for sequences. That is, we showed that there exist two summable, nowhere zero sequences whose convolution was identically zero.

This has been a brief survey of our accomplishments; more details can be found in the indicated publications. These accomplishments further our understanding of many aspects of estimation theory, of the performance of a contention resolution scheme for local area computer networks, of data compression for image processing, of signal detection theory, and of linear system theory.

REFERENCES

[1] G.L. Wise, "A note on a common misconception in estimation," *Systems and Control Letters*, Vol. 5, pp. 355-356, April 1985.

[2] R.S. Liptser and A.N. Shiryayev, *Statistics of Random Processes I.* New York: Springer-Verlag, 1977.

[3] R.S. Liptser and A.N. Shiryayev, *Statistics of Random Processes II.* New York: Springer-Verlag, 1978.

[4] I.I. Gihman and A.V. Skorohod, *The Theory of Stochastic Processes II.* New York: Springer-Verlag, 1975.

[5] I.I. Gihman and A.V. Skorohod, *The Theory of Stochastic Processes III.* New York: Springer-Verlag, 1979.

[6] R. Durrett, *Brownian Motion and Martingales in Analysis.* New York: Wiley, 1984, pp. 11-17.

[7] N. Ikeda and S. Watanabe, *Stochastic Differential Equations and Diffusion Processes.* Amsterdam: North Holland, 1981, p. 33.

[8] J. M. Morrison and G. L. Wise, "Continuity of filtrations of sigma algebras," *Statistics and Probability Letters*, Vol. 6, pp. 55-60, September 1987.

[9] J. M. Morrison and G. L. Wise, "The martingale convergence theorem, with tears," *Systems and Control Letters*, Vol. 9, pp. 275-279, September 1987.

[10] A. E. Wessel and G. L. Wise, "On estimation of random variables via the martingale convergence theorem," *Systems and Control Letters,* Vol. 11, pp. 61- 64, July 1988.

[11] J. M. Morrison and G. L. Wise, "Some comments on the convergence rate for adaptive schemes," *Proceedings of the Thirty-First Midwest Symposium on Circuits and Systems,* August 10-12, 1988, St. Louis, Missouri, pp. 973-975.

[12] S. Sherman, "Non–Mean–Square Error Criteria," *IRE Trans. Inform. Theory,* vol. IT–4, pp. 125–126, Sept. 1958.

[13] E. B. Hall and G. L. Wise, "Simultaneous optimal estimation over a family of fidelity criteria," *Proceedings of the 1987 Conference on Information Sciences and Systems*, Baltimore, Maryland, March 25-27, 1987, pp. 65-70.

[14] E. B. Hall and G. L. Wise, "On an aspect of nonlinear estimation," *Proceedings of the Twenty-Seventh Annual Allerton Conference on Communication, Control, and Computing,* Monticello, Illinois, September 27-29, 1989, pp. 692-701.

[15] E. B. Hall, A. E. Wessel, and G. L. Wise, "Some aspects of fusion in estimation theory," to appear in *IEEE Transactions on Information Theory.*

[16] E. B. Hall, A. E. Wessel, and G. L. Wise, "On fusion in estimation theory," *Proceedings of the Twenty-Sixth Annual Allerton Conference on Communication, Control, and Computing,* Monticello, Illinois, September 28-30, 1988, pp. 599-608.

[17] W. E. Wessel, E. B. Hall, and G. L. Wise, "Importance sampling via a simulacrum," to appear in *Journal of the Franklin Institute.*

[18] A. E. Wessel, E. B. Hall, and G. L. Wise, "Some comments on importance sampling," *Proceedings of the 1988 Conference on Information Sciences and Systems,* Princeton, New Jersey, March 16-18, 1988, pp. 325-329.

[19] G. L. Wise, "Counterexamples in detection and estimation," *Advances in Communications and Signal Processing* in Lecture Notes in Control and Information Sciences, pp. 121-132, edited by W. A. Porter and S. C. Kak, Springer-Verlag, New York, 1989.

[20] E. B. Hall and G. L. Wise, "Conditioning: a critical review," *Proceedings of the Twenty-Third Annual Conference on Information Sciences and Systems,* Baltimore, Maryland, March 22-24, 1989, pp. 271-276.

[21] E. B. Hall and G. L. Wise, "Some pathologies associated with maximum likelihood estimation," to appear in *Proceedings of the Twentieth-Fourth Annual Conference on Information Sciences and Systems,* Princeton, New Jersey, March 21-23, 1990.

[22] E. B. Hall and G. L. Wise, "Some comments on filtering in estimation theory," *Proceedings of the Thirty-Second Midwest Symposium on Circuits and Systems,* Urbana, Illinois, August 14-16, 1989, pp. 1162-1165.

[23]. G. L. Wise and E. B. Hall, "Just how good is the Kalman filter," to appear in *Proceedings of the Twentieth-Fourth Annual Conference on Information Sciences and Systems,* Princeton, New Jersey, March 21-23, 1990

[24] M.K. Molloy, "Collision resolution on the CSMA/CD bus," in *Proceedings Ninth Conf. Local Comput. Networks*, Minneapolis, MN, Oct. 8-10, 1984, pp. 44-47.

[25] Y.- C. Liu and G. L. Wise, "Performance of a CSMA/CD protocol for local area networks," *IEEE Journal on Selected Areas in Communications*, Vol. SAC-5, pp. 948-955, July 1987.

[26] N.C. Griswold, D.R. Halverson, and G.L. Wise, "A note on adaptive block truncation coding for image processing," to appear in *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-35, August, 1987.

[27] D. R. Halverson and G. L. Wise, "A note on an invariance property of signal detectors," *Computational Statistics and Data Analysis*, Vol. 6, pp. 59-64, January 1988.

[28] J. M. Morrison and G. L. Wise, "A functional analytic aspect of system theory," *Proceedings of the Twenty-Third Annual Conference on Information Sciences and Systems*, Baltimore, Maryland, March 22-24, 1989, pp. 289-293.

[29] E. B. Hall and G. L. Wise, "An algebraic aspect of linear system theory," to appear in *IEEE Transactions on Circuits and Systems.*

[30] J. M. Morrison and G. L. Wise, "A functional analytic aspect of system theory," *Proceedings of the Twenty-Third Annual Conference on Information Sciences and Systems,* Baltimore, Maryland, March 22-24, 1989, pp. 289-293.

[31] J. M. Morrison and G. L. Wise, "Some comments on the convergence rate for adaptive schemes," *Proceedings of the Thirty-First Midwest Symposium on Circuits and Systems,* August 10-12, 1988, St. Louis, Missouri, pp. 973-975.

[32] E. B. Hall and G. L. Wise, "A curiosum concerning discrete time convolution," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* Vol. 38, pp. 1058-1059, June 1990.

A copy of [20] is appended.

CONDITIONING: A CRITICAL REVIEW

Eric B. Hall
Department of Electrical and Computer Engineering
The University of Texas at Austin
Austin, Texas 78712

Gary L. Wise
Department of Electrical and Computer Engineering
and Department of Mathematics
The University of Texas at Austin
Austin, Texas 78712

ABSTRACT

The concept of conditioning in probability theory forms the basis for study in many areas of information sciences and systems. Even so, the subject of conditioning is often shrouded in heuristics, misunderstood, and misused. This paper considers several aspects of conditioning with an emphasis on applications and explores several consequences of an overly cavalier approach to the oft neglected measure-theoretic subtleties involved in this area.

I. INTRODUCTION

Conditioning in probability theory is a widely recurring concept in many areas of information sciences and systems. For example, conditioning is central to many popular techniques in applied probability and, in fact, lies at the heart of many aspects of estimation theory. In spite of this widespread popularity, the subject of conditioning is commonly misunderstood and tools involving conditioning are frequently misapplied. To rephrase Doob [5, p.v], conditioning is simply a branch of measure theory, and no attempt should be made to sugarcoat this fact. Unfortunately, many efforts at research have apparently been undertaken without appropriate concern for the measure-theoretic subtleties associated with the concept of conditioning. In this paper we review several aspects of conditioning and make a modest attempt to suggest caveats which seem to have been frequently overlooked by many in this area. Although several of our results are well known to the specialist in measure theory, they nevertheless seem to have been overlooked by many working in information sciences and systems.

In what follows, for a topological space T, we will let $\mathcal{B}(T)$ denote the family of Borel subsets of T. Also, we recall that a subset of a set is said to be cocountable if its complement is countable. Further, for a subset S, we will let S^c denote the complement of S, and we will let I_S denote the indicator function of S. In addition, we will let N denote the set of positive integers, Q denote the set of rational numbers, and R denote the set of real numbers. Finally, for a random variable X, $\sigma(X)$ will denote the σ-subalgebra generated by X.

II. SIGMA-ALGEBRAS

The topic of σ-algebras is basic to the subject of conditioning since conditioning is conventionally taken with respect to a σ-algebra. In many cases the σ-algebra of interest is that generated by some random variables representing data. Hence, in applications, it is common to treat σ-algebras as somehow representing knowledge or information associated with data. Consider the following example from [2, pp.458-459] which shows that associating σ-algebras with knowledge, or information, as commonly understood, can lead to incorrect conclusions.

Consider the probability space $([0, 1], \mathcal{B}([0, 1]), \lambda)$, where λ denotes Lebesgue measure on $\mathcal{B}([0, 1])$, and consider the σ-subalgebra $\mathcal{G}$ given by the family of all subsets of [0, 1] which are either countable or cocountable. Now, for $B \in \mathcal{B}([0, 1])$, consider the conditional probability $P(B \mid \mathcal{G})$. Since $\mathcal{G}$ contains all singletons $\{\omega\}$, and hence might be seen as being completely informative, an overly cavalier investigator might suppose that $P(B \mid \mathcal{G})$ is equal to I_B. In other words, one might rationalize that to know the sets in $\mathcal{G}$ implies that one knows ω itself and hence knows whether or not ω is contained in B, leading to the conclusion that $P(B \mid \mathcal{G})$ should be one when ω is contained in B and zero otherwise. It follows trivially, however, from the definition of conditional probability, that $P(B \mid \mathcal{G}) = P(B)$, except possibly off of a countable subset of [0, 1].

For another example, consider a probability space $(\Omega, \mathcal{F}, P)$. A commonly used model in estimation theory involves the model of data as a filtration $\{\mathcal{F}_n: n \in N\}$ of σ-subalgebras of $\mathcal{F}$. Suppose that the σ-algebra $\mathcal{F}$ is separable; that is, suppose $\mathcal{F}$ is generated by a countable family of subsets of Ω. Does it follow that $\mathcal{F}_n$ is separable for each n? As the following example illustrates, σ-subalgebras of separable σ-algebras need not be separable.

Assume that $\Omega = [0, 1]$ and $\mathcal{F} = \mathcal{B}([0, 1])$. Further, let $\mathcal{G}$ be the σ-subalgebra of $\mathcal{F}$ given by the countable and cocountable subsets of [0, 1]. Since $\mathcal{F} = \sigma((a, b): 0 \le a < b \le 1$ and $a,b \in Q)$ it follows that $\mathcal{F}$ is separable. Assume now that $\mathcal{G}$ is also separable; that is, assume that $\mathcal{G} = \sigma(A_n: n \in N)$ where A_n is a subset of [0, 1] for each n. Since $\mathcal{G}$ contains only countable and cocountable subsets of [0, 1], we may assume that A_n is countable for each n. Let $B = \bigcup_{n=1}^{\infty} A_n$, and note that B is also a countable subset of [0, 1]. Hence, there exists a real number x in [0, 1] which is not an element of B. Notice also that if $\mathcal{D}$ is the family of all subsets of B and their complements, then $\mathcal{D}$ is a σ-subalgebra such that $\mathcal{G} \supset \mathcal{D} \supset \sigma(A_n: n \in N)$. But, $\mathcal{D} \neq \mathcal{G}$ since $\{x\}$ is in $\mathcal{G}$ but not in $\mathcal{D}$. This contradiction implies that $\mathcal{G}$ is not separable even though it is a σ-subalgebra of the separable σ-algebra $\mathcal{F}$.

Now, let $(\Omega, \mathcal{F})$ be a measurable space, and let $\mathcal{P}$ be a family of probability measures on $(\Omega, \mathcal{F})$. The triple $(\Omega, \mathcal{F}, \mathcal{P})$ is called a probability structure. If $\mathcal{S}$ is a σ-sub-

Presented at the Twenty Third Annual Conference on Information Sciences and Systems, March 22-24, 1989, Baltimore, Maryland; to be published in the Proceedings of the conference.

algebra of $\mathcal{F}$, we say that $\mathcal{S}$ is sufficient if for each $\mathcal{F}$-measurable bounded real valued function f defined on Ω, there exists an $\mathcal{S}$-measurable bounded real valued function g defined on Ω such that $\int_A f\,dP = \int_A g\,dP$ for each A in $\mathcal{S}$ and for all P in $\mathcal{P}$. That is, g is almost surely [P] equal to the conditional expectation of f conditioned on $\mathcal{S}$ when P is the relevant probability measure. Note that although g does not depend on P, the set of P-measure zero might depend on P. It might be tempting and pleasing to the intuition to suppose that if $\mathcal{S}$ were a sufficient σ-subalgebra, then any σ-subalgebra of $\mathcal{F}$ which included $\mathcal{S}$ as a subset would also be sufficient. The following example from [3] constructs a nonsufficient σ-subalgebra which includes a sufficient σ-subalgebra.

Let $\mathcal{P}$ denote the family of probability measures P on $(\mathbf{R}, \mathcal{B}(\mathbf{R}))$ such that $P(B) = P(-B)$ for any set B in $\mathcal{B}(\mathbf{R})$ where, for any subset B of R, we define $-B = \{x \in \mathbf{R} : -x \in B\}$. Let $\mathcal{A} = \{B \in \mathcal{B}(\mathbf{R}) : B = -B\}$ and note that $\mathcal{A}$ is a σ-subalgebra of $\mathcal{B}(\mathbf{R})$. Further, $\mathcal{A}$ is a sufficient σ-subalgebra since, given any bounded Borel measurable function f, $g(x) = (f(x) + f(-x))/2$ is an $\mathcal{A}$-measurable function for which $\int_A f\,dP = \int_A g\,dP$ for any $A \in \mathcal{A}$ and any $P \in \mathcal{P}$.

Suppose now that Z is a subset of R which contains 0 and for which $Z = -Z$. Also, define $\mathcal{D} = \{B \cup A : B \in \mathcal{B}(\mathbf{R}), B \subset Z, \text{ and } A \in \mathcal{A}\}$. A straightforward examination shows that $\mathcal{D}$ is a σ-subalgebra of $\mathcal{B}(\mathbf{R})$ which includes $\mathcal{A}$.

Assume that $\mathcal{D}$ is a sufficient σ-subalgebra and let f be a bounded Borel measurable function. Then there exists a $\mathcal{D}$-measurable function g for which $\int_D f\,dP = \int_D g\,dP$ for any $D \in \mathcal{D}$ and any $P \in \mathcal{P}$. Let $x \in Z$ and note that $\{x\} \in \mathcal{D}$. Choosing $D = \{x\}$ above then implies that $f(x)P(\{x\}) = g(x)P(\{x\})$ for any measure P in $\mathcal{P}$. Now let $x \in Z^c$ and note that $\{x, -x\} \in \mathcal{D}$, $\{x\} \notin \mathcal{D}$, and $\{-x\} \notin \mathcal{D}$. Letting $D = \{x, -x\}$ above implies that $(f(x) + f(-x))P(\{x\}) = 2g(x)P(\{x\})$ since $P(\{x\}) = P(\{-x\})$, by definition of $\mathcal{P}$, and $g(x) = g(-x)$, since g is $\mathcal{D}$-measurable. Given any $x \in \mathbf{R}$, there exists a measure P in $\mathcal{P}$ for which $P(\{x\}) > 0$. Thus, we see that $g(x) = f(x)$ if $x \in Z$ and $g(x) = (f(x) + f(-x))/2$ if $x \in Z^c$. Let $f(x) = -1$ if $x < 0$ and $f(x) = 1$ if $x \geq 0$. This choice for f implies that g, as defined above, is nonzero on Z and zero on Z^c. Hence, we have that $Z = \{g^{-1}(\{0\})\}^c \in \mathcal{D}$. Now choose a subset Z_0 of R which contains 0, is such that $Z_0 = -Z_0$, and which is not an element of $\mathcal{B}(\mathbf{R})$. (Such sets abound.) Substituting Z_0 for Z thus implies, based on the above discussion, that $Z_0 \in \mathcal{D}$. But Z_0 cannot be in $\mathcal{D}$ since $Z_0 \notin \mathcal{B}(\mathbf{R})$. This contradiction implies that $\mathcal{D}$ is not a sufficient σ-subalgebra even though it includes the sufficient σ-subalgebra $\mathcal{A}$.

Filtrations of σ-algebras play a prominent role in many areas of conditioning. A common misconception concerning filtrations regards the relationship between the regularity of the sample paths of a random process and the continuity of its canonical filtration. In [9] examples are given in which a separable random process with a continuous filtration has nonmeasurable sample paths, a random process with infinitely differentiable sample paths has a discontinuous canonical filtration, and a random process taking values in [0, 1] has a canonical filtration which is everywhere discontinuous.

III. CONDITIONAL PROBABILITY

Consider a subset H of the interval [0, 1] with the properties that the outer Lebesgue measure of H is 1 and the inner Lebesgue measure of H is 0. (For a construction of such a set, the interested reader is referred to [8, pp.67-70].) Further, let $\Omega = [0, 1]$ and let λ denote Lebesgue measure on $\mathcal{B}([0, 1])$. Define $\mathcal{F} = \{(H \cap B_1) \cup (H^c \cap B_2) : B_1, B_2 \in \mathcal{B}([0, 1])\}$ and note that $\mathcal{F}$ is a σ-algebra on Ω and that $\mathcal{B}([0, 1])$ is a σ-subalgebra of $\mathcal{F}$. Now, define a probability measure $P:\mathcal{F} \to [0, 1]$ on the measurable space $(\Omega, \mathcal{F})$ via $P((H \cap B_1) \cup (H^c \cap B_2)) = (\lambda(B_1) + \lambda(B_2))/2$ to obtain a probability space $(\Omega, \mathcal{F}, P)$. (That P is well-defined follows from the properties of H.)

Consider now this probability space $(\Omega, \mathcal{F}, P)$. The following example, adapted from [2, p.464, 33.13], shows that conditional probabilities need not be measures.

Since $P(H) = 1/2$ and $P(B) = \lambda(B)$ for $B \in \mathcal{B}([0, 1])$ implies that that $P(H \cap B) = \lambda(B)/2 = P(H)P(B)$, it follows that H is independent of $\mathcal{B}([0, 1])$. Let F be a set in $\mathcal{F}$ with probability zero and assume that $P(\cdot \mid \mathcal{B}([0, 1]))(\omega)$ is a probability measure on $\mathcal{F}$ for each ω outside of the null set F. Note that there exists a collection $\{A_n : n \in \mathbf{N}\}$ of subsets of [0, 1] such that $\mathcal{B}([0, 1]) = \sigma(\{A_n : n \in \mathbf{N}\})$ and such that $\{A_n : n \in \mathbf{N}\}$ is closed under finite intersections. Define $K_n = \{\omega \in \Omega : P(A_n \mid \mathcal{B}([0, 1]))(\omega) = I_{A_n}(\omega)\}$ and note that $K_n \in \mathcal{B}([0, 1])$ and $P(K_n) = 1$ for all $n \in \mathbf{N}$ since $P(A_n \mid \mathcal{B}([0, 1])) = I_{A_n}$ a.s. Now, let $K = \bigcap_{n=1}^{\infty} K_n \cap F^c$ and note that $P(K) = 1$. Further, note that the function which, for a fixed ω in K, maps an element B of $\mathcal{B}([0, 1])$ to $I_B(\omega)$ is a probability measure on $\mathcal{B}([0, 1])$ which agrees with $P(B \mid \mathcal{B}([0, 1]))(\omega)$ whenever $B \in \{A_n : n \in \mathbf{N}\}$. Thus, the Dynkin system theorem [1, p.169] implies that for $\omega \in K$, $P(B \mid \mathcal{B}([0, 1]))(\omega)$ is uniquely determined to be $I_B(\omega)$ for any set B in $\mathcal{B}([0, 1])$. Thus, in particular, if $\omega \in K$ then $P(\{\omega\} \mid \mathcal{B}([0, 1]))(\omega) = 1$. Now, recalling we assumed that

$P(\cdot \mid \mathcal{B}([0, 1]))(\omega)$ is a probability measure on $\mathcal{F}$ for each ω outside of the null set F, we see that if $\omega \in H \cap K$ then $P(H \mid \mathcal{B}([0, 1]))(\omega) \geq P(\{\omega\} \mid \mathcal{B}([0, 1]))(\omega) = 1$, and if $\omega \in H^c \cap K$ then $P(H \mid \mathcal{B}([0, 1]))(\omega) \leq P(\{\omega\}^c \mid \mathcal{B}([0, 1]))(\omega) = 0$. Thus, if $\omega \in K$, then $P(H \mid \mathcal{B}([0, 1]))(\omega) = I_H(\omega)$. But H and $\mathcal{B}([0, 1])$ are independent, and hence $P(H \mid \mathcal{B}([0, 1])) = P(H) = 1/2$ a.s. This contradiction implies that $P(\cdot \mid \mathcal{B}([0, 1]))(\omega)$ is not almost surely a probability measure on $\mathcal{F}$. Hence, a conditional probability is not necessarily a measure.

A regular conditional probability allows one to sidestep many of the undesirable aspects of conditional probability since a regular conditional probability is by definition required to be a measure for each fixed $\omega \in \Omega$. Unfortunately, however, regular conditional probabilities do not always exist. In fact, the situation detailed above, in addition to showing that a conditional probability need not be a measure, also provides an example in which a regular conditional probability does not exist.

IV. CONDITIONAL INDEPENDENCE

The concept of conditional independence arises frequently in many aspects of probability theory. For example, the concept plays an important role in the study of Markov processes. Unfortunately, misconceptions often arise regarding the relationship between conditional independence and independence. As the following examples adapted (with a correction) from [4, p.221] indicate, the notions of independence and conditional independence taken with respect to a nontrivial σ-subalgebra are unrelated.

Consider a probability space $(\Omega, \mathcal{F}, P)$ and a σ-subalgebra $\mathcal{H}$ of $\mathcal{F}$. Further, let $\mathcal{H}_1$ and $\mathcal{H}_2$ be two families each composed of elements from $\mathcal{F}$. The families $\mathcal{H}_1$ and $\mathcal{H}_2$ are said to be conditionally independent given $\mathcal{H}$ if $P(A_1 \cap A_2 \mid \mathcal{H}) = P(A_1 \mid \mathcal{H})\, P(A_2 \mid \mathcal{H})$ a.s. for all $A_1 \in \mathcal{H}_1$ and $A_2 \in \mathcal{H}_2$. Further, two random variables X and Y defined on $(\Omega, \mathcal{F}, P)$ are said to be conditionally independent given $\mathcal{H}$ if $\sigma(X)$ and $\sigma(Y)$ are conditionally independent given $\mathcal{H}$.

Let X_1 and X_2 be two independent identically distributed random variables such that $P(X_1 = 1) = P(X_1 = -1) = 1/2$. Further, let $Z = X_1 + X_2$, and let $A_i = X_i^{-1}(\{1\})$ for i = 1 and 2. In this case, $P(A_i \mid Z) = 1/2$ on $Z^{-1}(\{0\})$ for i = 1 or 2, and $P(A_1 \cap A_2 \mid Z) = 0$ on $Z^{-1}(\{0\})$. In particular, $P(A_1 \cap A_2 \mid Z) \neq P(A_1 \mid Z)\, P(A_2 \mid Z)$ on an event of positive probability. Thus, the independent random variables X_1 and X_2 are not conditionally independent given $\sigma(Z)$.

Consider now three mutually independent random variables Y_1, Y_2, and Y_3 such that each random variable takes on only integer values and such that $P(Y_i = m) < 1$ for all integers m and for i = 1, 2, 3. Further, let $S_2 = Y_1 + Y_2$ and $S_3 = Y_1 + Y_2 + Y_3$ and notice that Y_1 and S_3 are dependent random variables. Let $B_i = S_2^{-1}(\{i\})$ for each integer i. There exists k such that B_k has positive probability. On such a set B_k we have that

$$\begin{aligned}
P(Y_1 = i, S_3 = j \mid S_2) &= P(Y_1 = i, S_2 = k, S_3 = j) / P(S_2 = k) \\
&= P(Y_1 = i)\, P(Y_2 = k - i)\, P(Y_3 = j - k) / P(S_2 = k) \\
&= (P(Y_1 = i)\, P(Y_2 = k - i) / P(S_2 = k))\, P(Y_3 = j - k) \\
&= (P(Y_1 = i, S_2 = k) / P(S_2 = k))\, P(Y_3 = j - k) \\
&= P(Y_1 = i \mid S_2)\, P(Y_3 = j - k) \\
&= P(Y_1 = i \mid S_2)\, (P(Y_3 = j - k)\, P(S_2 = k) / P(S_2 = k)) \\
&= P(Y_1 = i \mid S_2)\, (P(S_2 = k, S_3 = j) / P(S_2 = k)) \\
&= P(Y_1 = i \mid S_2)\, P(S_3 = j \mid S_2).
\end{aligned}$$

Thus, we conclude that even though Y_1 and S_3 are dependent random variables, Y_1 and S_3 are conditionally independent given $\sigma(S_2)$.

V. CONDITIONAL EXPECTATION

Let X be an integrable random variable defined on the probability space $(\Omega, \mathcal{F}, P)$, and let $\mathcal{H}$ be a σ-subalgebra of $\mathcal{F}$. Then can the conditional expectation $E[X \mid \mathcal{H}]$ be expressed as $\int_\Omega X\, dP(\cdot \mid \mathcal{H})$, where $P(\cdot \mid \mathcal{H})$ denotes conditional probability given $\mathcal{H}$? The alert reader will immediately give a negative response to this question, since, recalling Section III, $P(\cdot \mid \mathcal{H})$ might not be a measure and hence the preceding integral might not even be defined.

The following example counters a common misconception concerning versions of conditional expectations. In particular, a random variable is given which is equal a.s. to a conditional expectation yet is not a version of the conditional expectation.

Consider the probability space consisting of [0, 1], $\mathcal{B}([0, 1])$, and Lebesgue measure on $\mathcal{B}([0, 1])$, and let $\mathcal{G}$ denote the σ-algebra consisting of the countable and cocountable subsets of [0, 1]. Let X be the identity map on [0, 1] and note that $E[X \mid \mathcal{G}] = 1/2$ a.s. Further, let $Y = \frac{1}{2}(1 - I_C)$ where C denotes the Cantor ternary set. Note that $Y = E[X \mid \mathcal{G}]$ a.s., yet Y is not $\mathcal{G}$-measurable (since C is neither countable nor cocountable) and hence is not a version of $E[X \mid \mathcal{G}]$.

Another commonly occurring misconception regarding conditional expectation is that it is a "smoothing" operator. Consider, for example, a random process $\{X(t) : t \in \mathbf{R}\}$ defined on a probability space $(\Omega, \mathcal{F}, P)$ and a σ-subalgebra $\mathcal{H}$ of $\mathcal{F}$. It has been argued by some (see for instance several recent papers in the area of perturbation analysis) that $E[X(t) \mid \mathcal{H}]$ is "smoother" than X(t) as a function of t. To dispel this absurd notion simply let X(t) be an $\mathcal{H}$-measurable random process which is discontinuous everywhere; the version of $E[X(t) \mid \mathcal{H}]$ given by X(t) obviously retains this same property.

Perhaps a little less obvious is the fact that, for a random variable X on $(\Omega, \mathcal{F}, P)$ and a σ-subalgebra $\mathcal{G}$ of $\mathcal{F}$, $E[X \mid \mathcal{G}]$ need not be as "smooth" a function of ω as X. Consider for instance the probability space given by the interval [0, 1], the σ-algebra $\mathcal{G}$ given by the countable and cocountable subsets of [0, 1], and Lebesgue measure on $\mathcal{G}$. If we let $X = 1$, then a version of $E[X \mid \mathcal{G}]$ is given by $1 - I_B$ where B equals the set of rationals in [0, 1]. Hence, even though X is everywhere continuous, there exists a version of $E[X \mid \mathcal{G}]$ which is everywhere discontinuous.

A commonly encountered property of conditional expectation is the so-called nesting property. Unfortunately, this property is sometimes misapplied. In this example, from [6], it is shown that $E[E[X \mid Y]]$ may exist even when the expectation of X does not exist. In other words, before calculating $E[E[X \mid Y]]$ and claiming one has found the mean of X, it is necessary to first ascertain that the mean of X actually exists.

Consider random variables X and Y defined on the same probability space such that Y possesses a probability density function g(y) given by $g(y) = \frac{1}{\sqrt{2\pi y}} \exp\left(\frac{-y}{2}\right)$; $y > 0$, and, for each $y > 0$, Y is such that a conditional density function of X given $Y = y$, denoted by $f(x|y)$, exists and is given by $f(x|y) = \sqrt{\frac{y}{2\pi}} \exp\left(\frac{-yx^2}{2}\right)$; $y > 0$ and $x \in \mathbf{R}$. It follows immediately that $E[X \mid Y] = 0$ a.s. and therefore $E[E[X \mid Y]] = 0$. Notice, however, that the mean of X does not exist since X has a Cauchy density h(x) given by h(x) =

$$\int_{\mathbf{R}} f(x|y)\, g(y)\, dy = \frac{1}{\pi(1 + x^2)} \quad \text{for } x \in \mathbf{R}.$$

For another example, consider random variables X and Y each defined on the same probability space $(\Omega, \mathcal{F}, P)$ and a σ-subalgebra $\mathcal{M}$ of $\mathcal{F}$. Another commonly encountered misconception concerning conditional expectation is that $E[X \mid \mathcal{M}]$ and $E[Y \mid \mathcal{M}]$ are independent if X and Y are independent. The following counterexample, which [12, p.133] attributes to C. Sugahara, demonstrates that in general this conclusion is false.

Let U and V be independent random variables, each defined on the probability space $(\Omega, \mathcal{F}, P)$, and each having a zero mean, unit variance Gaussian distribution. Define $X = U + V$ and $Y = U - V$, and note that X and Y are independent random variables each having a zero mean Gaussian distribution with a variance of 2. Further,
$E[X \mid \sigma(U)] = E[U + V \mid \sigma(U)] = U + E[V \mid \sigma(U)] = U + E[V] = U$ a.s., and $E[Y \mid \sigma(U)] = E[U - V \mid \sigma(U)] = U - E[V \mid \sigma(U)] = U - E[V] = U$ a.s. Hence, any version of $E[X \mid \sigma(U)]$ and any version of $E[Y \mid \sigma(U)]$ are equal almost surely to the same positive variance Gaussian random variable and hence cannot be independent. Further, we note that even the ubiquitous Gaussian assumption does not alleviate this problem.

Fatou's lemma and uniform integrability are powerful tools in analysis and are often relied upon in the area of estimation theory. We recall that if a sequence of random variables is uniformly integrable then almost sure convergence implies convergence of the corresponding expectations. Convergence of conditional expectations with respect to an arbitrary σ-subalgebra, however, does not follow in general. The following example, adapted from [16], describes a situation in which Fatou's lemma does not hold and in which uniform integrability and almost sure convergence do not imply that the corresponding conditional expectations converge.

Let $\Omega = (0, 1) \times (0, 1)$, let $\mathcal{H} = \{B \times (0, 1): B \in \mathcal{B}((0, 1))\}$, and note that $\mathcal{H}$ is a σ-algebra on Ω. Let μ denote Lebesgue measure on $\mathcal{B}((0, 1))$, and let P denote Lebesgue measure on $\mathcal{B}((0, 1) \times (0, 1))$. For each positive integer n, let $B_n = (0, \frac{1}{n})$, and let A_n denote the n-th term in the sequence $(0, \frac{1}{2}), (\frac{1}{2}, 1), (0, \frac{1}{4}), (\frac{1}{4}, \frac{1}{2}), (\frac{1}{2}, \frac{3}{4}), (\frac{3}{4}, 1), (0, \frac{1}{8}), (\frac{1}{8}, \frac{1}{4}), \cdots$. Note that $(\Omega, \mathcal{B}((0, 1) \times (0, 1)), P)$ is a probability space, and that $\mathcal{H}$ is a σ-subalgebra of $\mathcal{B}((0, 1) \times (0, 1))$. Now, for each positive integer n, define a random variable $X_n(x, y) = \frac{1}{\mu(B_n)} I_{A_n \times B_n}(x, y)$. Let $B \in \mathcal{B}((0, 1))$, and note that

$$\int_{B\times(0,1)} X_n\, dP = \int_{B\times(0,1)} \frac{1}{\mu(B_n)} I_{A_n \times B_n}\, dP$$

$$= \mu(A_n \cap B) = \int_{B\times(0,1)} I_{A_n \times (0,1)}\, dP,$$ which thus implies that $E[X_n \mid \mathcal{H}] = I_{A_n \times (0,1)}$ a.s. Now, note that $X_n \geq 0$ for each positive integer n, and that $X_n \to 0$ as $n \to \infty$. Note also that, since $E[|X_n|] = \mu(A_n) \to 0$, the random variables $\{X_n: n \in \mathbf{N}\}$ are uniformly integrable. Further, note that $\overline{\lim}_{n\to\infty} I_{A_n \times (0,1)} = 1$ and that $\underline{\lim}_{n\to\infty} I_{A_n \times (0,1)} = 0$. Thus, we see that, even though the random variables $\{X_n: n \in \mathbf{N}\}$ are uniformly integrable, $\overline{\lim}_{n\to\infty} E[X_n \mid \mathcal{H}] = 1$ a.s. and $\underline{\lim}_{n\to\infty} E[X_n \mid \mathcal{H}] = 0$ a.s. In particular, Fatou's lemma does not hold and the conditional expectations do not converge.

VI. REGRESSION FUNCTIONS

Given two random variables X and Y defined on the same probability space, a common problem concerns the determination of the form of the regression function $E[X \mid Y=y]$. For example, [13] considers this problem when both X and Y are uniformly distributed. In this example, we show that the existence of a joint probability density function for X and Y in no way guarantees that the regression function will obey any regularity property, other than Borel measurability.

Let $g: \mathbf{R} \to \mathbf{R}$ be Borel measurable and define $f(x,y) = \frac{1}{4}\exp(-\exp(|y|)\,|x - g(y)|)$. Note that f(x,y) is a joint probability density function since $\int_{\mathbf{R}} \int_{\mathbf{R}} f(x,y)\, dx\, dy$

$$= \int_{\mathbf{R}} \int_{\mathbf{R}} \frac{1}{4} \exp(-\exp(|y|)\,|x - g(y)|)\, dx\, dy$$

$$= \int_R \int_R \frac{1}{4} \exp(-\exp(|y|)\,|z|)\,dz\,dy = \int_R \frac{1}{2} \exp(-|y|)\,dy = 1.$$

Let X and Y be random variables such that the pair (X, Y) has a joint density function given by f(x, y). Notice from the above calculation that a second marginal density of f(x, y) is given by $f_Y(y) = \frac{1}{2}\exp(-|y|)$. Recalling that

$E[X \mid Y=y] = \int_R x \frac{f(x,y)}{f_Y(y)}\,dx$ and substituting for $f_Y(y)$ implies that $E[X \mid Y=y]$

$$= 2\exp(|y|) \int_R \frac{x}{4} \exp(-\exp(|y|)\,|x - g(y)|)\,dx$$

$$= 2\exp(|y|) \int_R ((z + g(y))\,\frac{1}{4} \exp(-\exp(|y|)\,|z|)\,dz$$

$$= 2\exp(|y|)\,g(y) \frac{1}{2\exp(|y|)} = g(y).$$

Hence, the random variables X and Y with the joint density function f(x, y) are such that $E[X \mid Y=y] = g(y)$ where we recall that $g(\cdot)$ was an arbitrarily selected Borel measurable function.

VII. MEAN SQUARE ESTIMATION

One of the most common misconceptions in estimation theory is that conditional expectation minimizes mean square error. This mistaken concept arises in estimation and filtering applications in engineering as well as in many L_2 minimization problems in probability and statistics. As the following example from [15] indicates, even for bounded random variables, conditional expectation may not even come close to minimizing the mean square error even though there exists a function mapping the reals into the reals by which the random variable of interest may be estimated precisely.

Consider the set H and the probability space $(\Omega, \mathcal{F}, P)$ used in Section III. Let λ denote Lebesgue measure on $\mathcal{B}([0, 1])$. Further, let A be a fixed nonzero real number and define two random variables X and Y on $(\Omega, \mathcal{F}, P)$ via $X(\omega) = \omega$ and $Y(\omega) = A\,I_H(\omega)$. Notice that $\sigma(X) = \mathcal{B}([0, 1])$ and that $\sigma(Y) = \{\Omega, \emptyset, H, H^c\}$. Further, since $P(H) = 1/2$ and $P(B) = \lambda(B)$ for $B \in \mathcal{B}([0, 1])$, we see that $P(H \cap B) = \lambda(B)/2 = P(H)P(B)$, or that X is independent of Y. Hence $E[Y \mid X] = E[Y] = A/2$ a.s. which implies that $E[(Y - E[Y \mid X])^2] = E[(Y - A/2)^2] = A^2/4$. But, $Y(\omega) = A\,I_H(X(\omega))$ for all $\omega \in \Omega$. Thus, $E[(Y - A\,I_H(X))^2] = 0$. In other words, for this example there exists a function $f: \mathbf{R} \to \mathbf{R}$ such that $Y(\omega) = f(X(\omega))$ for all $\omega \in \Omega$ yet, by choice of A, $E[(Y - E[Y \mid X])^2]$ could be arbitrarily large. We note further that in this case $\sigma(Y)$ is finite, $\sigma(X)$ contains all singletons, and all moments of X and Y exist.

VIII. DISTRIBUTED ESTIMATION

Consider a random variable X and a set of random variables $\{Y_1, \ldots, Y_n\}$ all defined on the same probability space. A commonly considered problem in the area of distributed estimation is that of how to best fuse or combine estimates of the form $E[X \mid \mathcal{D}]$, where $\mathcal{D}$ is a nonempty proper subset of $\{Y_1, \ldots, Y_n\}$, in order to obtain a single good estimate of $E[X \mid Y_1, \ldots, Y_n]$. In the following example, from [7], a situation is described, using common distributions, in which any such method of fusion is useless.

For a positive integer n greater than one, consider a set of random variables $\{X, Y_1, \ldots, Y_n\}$ with a joint probability density function given, as in [11], by

$$f(x, y_1, \ldots, y_n) = \left(\frac{1}{\sqrt{2\pi}}\right)^{n+1} \exp\left[\frac{-1}{2}\left(x^2 + \sum_{i=1}^{n} y_i^2\right)\right]$$

$$\cdot \left[1 + x\exp\left(\frac{-x^2}{2}\right) \prod_{i=1}^{n} \left(y_i \exp\left(\frac{-y_i^2}{2}\right)\right)\right].$$ It follows straightforwardly that the set $\{X, Y_1, \ldots, Y_n\}$ is not mutually Gaussian and not mutually independent, yet any proper subset of $\{X, Y_1, \ldots, Y_n\}$ containing at least two random variables is mutually independent, mutually Gaussian, and identically distributed with each random variable having zero mean and unit variance. For any nonempty proper subset $\mathcal{D}$ of $\{Y_1, \ldots, Y_n\}$, we note that $E[X \mid \mathcal{D}] = 0$ a.s. since X is independent of $\mathcal{D}$. However, that $E[X \mid Y_1, \ldots, Y_n] = \frac{1}{2\sqrt{2}} Y_1 \cdots Y_n \exp\left[\frac{-1}{2}(Y_1^2 + Y_2^2 + \ldots + Y_n^2)\right]$ a.s. follows easily. Thus, since any Borel measurable function of the estimates $E[X \mid \mathcal{D}]$ where $\mathcal{D}$ ranges over all nonempty proper subsets of $\{Y_1, \ldots, Y_n\}$ would be constant almost surely, it would be absurd to attempt to estimate $E[X \mid Y_1, \ldots, Y_n]$ based on a *combination of these estimates*. Once again, notice that the oft used and much abused Gaussian assumption does not alleviate this difficulty.

IX. MARTINGALES

The subject of martingale theory is an important aspect of conditioning which finds many applications in information sciences and systems. The following example shows that a martingale may have a constant positive mean, converge a.s. to zero in finite time, and yet with positive probability exceed any real number.

Let $\{X_n: n \in \mathbf{N}\}$ be a sequence of mutually independent identically distributed random variables such that $P(X_1 = 0) = P(X_1 = 2) = 1/2$. Now, for each positive integer n, define $Y_n = X_1 X_2 \cdots X_n$, and note that $\{Y_n: n \in \mathbf{N}\}$ is a martingale and that $E[Y_n] = 1$ for all $n \in \mathbf{N}$. Further, notice that not only does the sequence $\{Y_n: n \in \mathbf{N}\}$ converge almost surely to zero, but with probability one, *only a finite* number of terms of the sequence $\{Y_n: n \in \mathbf{N}\}$ are nonzero. Even so, it follows easily that Y_n exceeds any real value with positive probability since $P(Y_n = 2^n) > 0$ for all $n \in \mathbf{N}$.

Consider now the following example from [14] which illustrates a pathology concerning the martingale convergence theorem. In particular, it shows that in certain circumstances the martingale convergence theorem might be useless as an estimation technique.

Consider the probability space $(\mathbf{R}, \mathcal{B}(\mathbf{R}), P)$ where P denotes zero mean, unit variance Gaussian measure on $(\mathbf{R}, \mathcal{B}(\mathbf{R}))$. Let P_* denote the inner P measure on $(\mathbf{R}, \mathcal{B}(\mathbf{R}))$. Let S be a subset of R such that $P_*(S) = P_*(S^c) = 0$. (That such sets exist is shown in [14].) Further, let $\mathcal{W} = \{(S \cap B_1) \cup (S^c \cap B_2): B_1, B_2 \in \mathcal{B}(\mathbf{R})\}$ and note that $\mathcal{W}$ is a σ-algebra on R which includes $\mathcal{B}(\mathbf{R})$. Define a probability measure μ on $(\mathbf{R}, \mathcal{W})$ via $\mu((S \cap B_1) \cup (S^c \cap B_2)) = (P(B_1)+P(B_2))/2$. (That μ is well-defined follows from the properties of S.) Note that the restriction of μ to $\mathcal{B}(\mathbf{R})$ is P.

Consider now the probability space $(\mathbf{R}, \mathcal{W}, \mu)$. Note first that S and S^c are each independent of $\mathcal{B}(\mathbf{R})$ since, for any Borel set B, $\mu(S \cap B) = \mu(B)/2 = \mu(S)\,\mu(B)$ and $\mu(S^c \cap B) = \mu(B)/2 = \mu(S^c)\,\mu(B)$. Now define a random variable X on $(\mathbf{R}, \mathcal{W}, \mu)$ via $X(x) = x\, I_S(x) - x\, I_{S^c}(x)$ and notice that, for any Borel set B, $\mu(X \in B) = \mu((S \cap B) \cup (S^c \cap \{x \in \mathbf{R}: -x \in B\})) = P(B)$ since P is symmetric. Hence, X is a Gaussian random variable with zero mean and unit variance. Further, note that $E[X(x) \mid \mathcal{B}(\mathbf{R})] = x\, E[I_S(x) - I_{S^c}(x) \mid \mathcal{B}(\mathbf{R})] = x\, E[I_S(x) - I_{S^c}(x)] = 0$ a.s. since the identity map is Borel measurable, S and S^c are independent of $\mathcal{B}(\mathbf{R})$, and $P(S) = P(S^c) = 1/2$.

Now, let $\{Y_k: k \in \mathbf{N}\}$ be any sequence of Borel measurable functions mapping R into R. Note that $\{Y_k: k \in \mathbf{N}\}$ is a sequence of random variables on $(\mathbf{R}, \mathcal{W}, \mu)$. Consider the martingale $\{X_k = E[X \mid Y_1, \ldots, Y_k] : k \in \mathbf{N}\}$. Since, given any $k \in \mathbf{N}$, $\mathcal{B}(\mathbf{R})$ includes $\sigma(Y_1, \ldots, Y_k)$, it follows that $X_k = E[X \mid Y_1, \ldots, Y_k] = E[\, E[X \mid \mathcal{B}(\mathbf{R})] \mid Y_1, \ldots, Y_k] = 0$ a.s. using the previous result. Hence, the martingale convergence theorem is completely useless in estimating the random variable X in terms of the random variables $\{Y_k: k \in \mathbf{N}\}$. Furthermore, note that for any sequence $\{s_k: k \in \mathbf{N}\}$ of positive real numbers, we could let $Y_k(x) = s_k\, x$. In this case, the above phenomena is exhibited when all of the random variables of concern are Gaussian. Finally, we note that, yet again, the ubiquitous Gaussian assumption does not protect us from this disturbing problem.

Another disturbing result concerning the martingale convergence theorem is detailed in [10]. There it is shown that the convergence rate guaranteed by the martingale convergence theorem can be arbitrarily slow. This result contrasts with the previous example in which the convergence was instantaneous, yet to the wrong random variable.

X. CONCLUSION

We hope these comments will be helpful to those using conditioning as a tool in investigations. Although some of these examples are undoubtedly well known to the specialist in measure theory, as previously mentioned, our experience indicates that these caveats have been overlooked by many working in the area of information sciences and systems. In conclusion, if this paper serves no other purpose, we hope it will serve as a reminder that conditioning can be a dangerous tool in the hands of amateurs.

ACKNOWLEDGEMENT

This research was supported by the Air Force Office of Scientific Research under Grant AFOSR-86-0026.

REFERENCES

1. Ash, R., *Real Analysis and Probability*, New York: Academic Press, 1972.
2. Billingsley, P., *Probability and Measure*, 2nd ed., New York: Wiley, 1986.
3. Burkholder, D. L., "Sufficiency in the Undominated Case," *The Annals of Mathematical Statistics*, Vol. 32, No. 4, pp. 1191-1200, Dec. 1961.
4. Chow, Y. S. and H. Teicher, *Probability Theory: Independence, Interchangeability, Martingales*, 2nd ed., New York: Springer-Verlag, 1988.
5. Doob, J. L., *Stochastic Processes*, New York: Wiley, 1953.
6. Enis, P., "On the Relation $E(X) = E\{E(X \mid Y)\}$," *Biometrika*, Vol. 60, pp. 432-433, 1973.
7. Hall, E. B., A. E. Wessel and G. L. Wise, "On Fusion in Estimation Theory," *Proceedings of the Twenty-Sixth Annual Allerton Conference on Communication, Control, and Computing*, Allerton, Illinois, Sept. 28-30, 1988, pp. 599-608.
8. Halmos, P. R., *Measure Theory*, New York: Springer-Verlag, 1974.
9. Morrison, J. M. and G. L. Wise, "Continuity of Filtrations of Sigma Algebras," *Statistics and Probability Letters*, Vol. 6, No. 1, pp. 55-60, Sept. 1987.
10. Morrison, J. M. and G. L. Wise, "The Martingale Convergence Theorem, With Tears," *Systems and Control Letters*, Vol. 9, No. 3, pp. 275-279, Sept. 1987.
11. Pierce, D. A. and R. L. Dykstra, "Independence and the Normal Distribution," *The American Statistician*, Vol. 23, No. 4, p. 39, Oct. 1969.
12. Romano, J. P. and A. F. Siegel, *Counterexamples in Probability and Statistics*, Monterey, California: Wadsworth and Brooks, 1986.
13. Vitale, A. R. and A. C. Pipkin, "Conditions on the Regression Function When Both Variables are Uniformly Distributed," *The Annals of Probability*, Vol. 4, No. 5, pp. 869-873, Oct. 1976.
14. Wessel, A. E. and G. L. Wise, "On Estimation of Random Variables via the Martingale Convergence Theorem," *Systems and Control Letters*, Vol. 11, No. 1, pp. 61-64, July 1988.
15. Wise, G. L., "A Note on a Common Misconception in Estimation," *Systems and Control Letters*, Vol. 5, pp. 355-356, April 1985.
16. Zheng, W., "A Note on the Convergence of Sequences of Conditional Expectations of Random Variables," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, Vol. 53, pp. 291-292, Sept. 1980.